Explanation in the Era of LLMs

NAACL 2024 tutorial
Section 4: Transformer Understanding

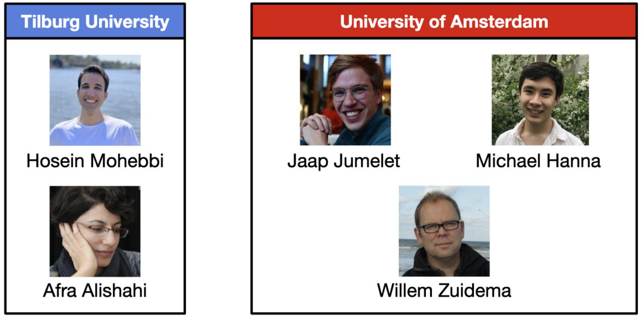
Presented by Sarah Wiegreffe
Thanks Xi and Chenhao for help with slides.

Tutorial @ EACL 2024: Transformer-Specific Interpretability

[Website](#)

[Slides](#)

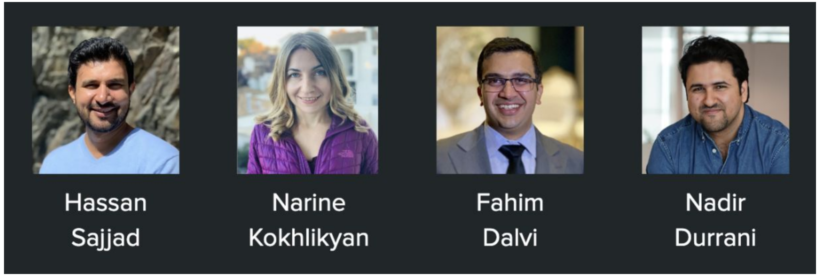
[Recording](#)



Tutorial @ NAACL 2021: Fine-grained Interpretation and Causation Analysis in Deep NLP Models

[Website/Slides](#)

[Recording](#)



Tutorial @ EMNLP 2022: Causal Inference for NLP

[Slides](#)/[Slides](#)

[Recording](#)



Mechanistic Interpretability Workshop @ ICML 2024

[Website](#)

Fazl Barez
Research Fellow University of Oxford

Mor Geva
Ass. Prof Tel Aviv University, Visiting Researcher Google Research

Lawrence Chan
PhD student UC Berkeley

Kayo Yin
PhD student UC Berkeley

Neel Nanda
Research Engineer Google DeepMind

Max Tegmark
Professor MIT

Survey Papers

[A Primer on the Inner Workings of Transformer Language Models](#)

[Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks](#)

Outline

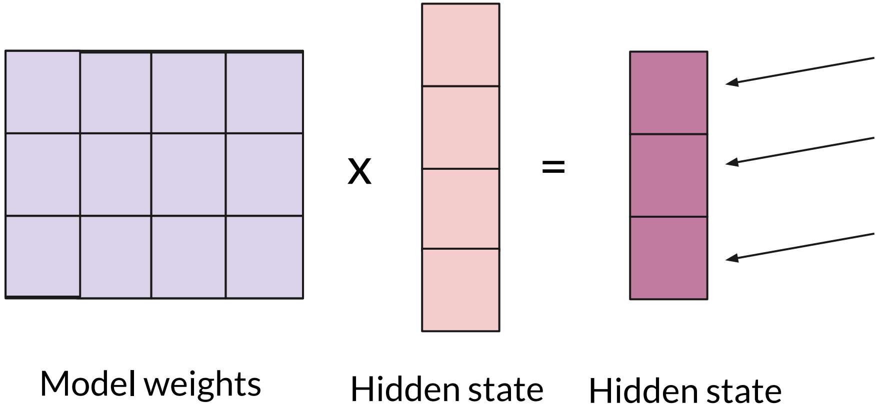
1. Neuron-level interpretability
    a. Sparse Autoencoders
2. Causal Mediation
    a. Activation Patching & variants
    b. Causal abstraction & other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

Neuron-Level Interpretability

Background

Interpreting Neurons

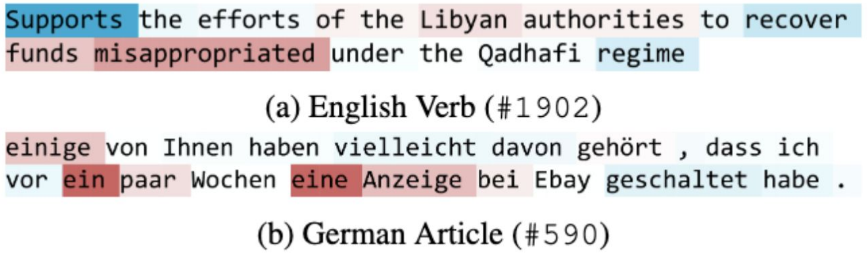
- Neuron = a single dimension of a hidden state representation
- Line of work traditionally does not consider structure



What pattern in the inputs will fire a neuron (i.e., cause high values at a particular dimension)?

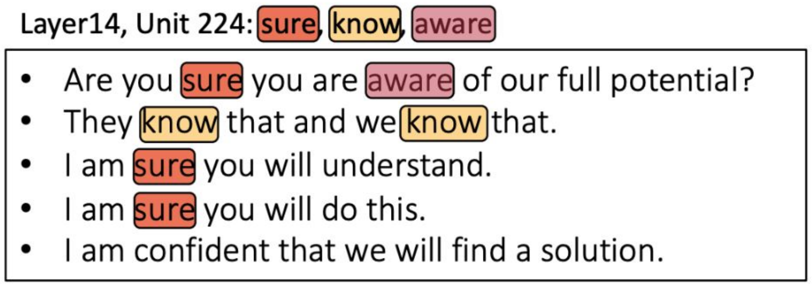
Interpreting Neurons of NLP Models

Activations of neurons for certain properties



(a) English Verb (#1902)

(b) German Article (#590)



[Bau et al., ICLR 2019] [Dalvi et al., AAAI 2019]

[Mu & Andreas, NeurIPS 2020] [Na et al., ICLR 2019]

Pitfalls of Neuron-Level Analysis in NLP

- Methods generally ignore interactions between neurons.
- There are a LOT of neurons in modern models.

[Sajjad et al. 2022, Antverg & Belinkov, ICLR 2022]

Pitfalls of Visualization/Looking at Examples

- Humans are biased towards simple and clear concepts.

> - *”What is the meaning behind the song ””Angel”” by Eric Clapton?”*
> - *”What’s the meaning of Johnny Cash’s song ””King of the Hill””?”*
> - *”What is the meaning behind the Tears for Fears song ””Mad World””, such as the lyric, ””All around me are familiar faces””?”*

Song titles? Syntactic sentence structure?

[Bolukbasi et al 2021, Olah et al 2020]

Pitfalls of Visualization/Looking at Examples

- Humans are biased towards simple and clear concepts.

- *”What is the meaning behind the song ””Angel”” by Eric Clapton?”*
- *”What’s the meaning of Johnny Cash’s song ””King of the Hill””?”*
- *”What is the meaning behind the Tears for Fears song ””Mad World””, such as the lyric, ””All around me are familiar faces””?”*

Song titles? Syntactic sentence structure?

- *On 16 June 2006, it was announced that Everton had entered into talks with Knowsley Council and Tesco over the possibility of building a new 55,000 seat stadium, ex-pandable to over 60,000, in Kirkby.*
- *On 15 September 1940, known as the Battle of Britain Day, an RAF pilot, Ray Holmes of No. 504 Squadron RAF rammed a German bomber he believed was going to bomb the Palace.*
- *On 20 August 2010, Queen’s manager Jim Beach put out a Newsletter stating that the band had signed a new contract with Universal Music.*

Historical events? Sentences with dates at the beginning?

- **Polysemanticity**: neurons “respond to multiple unrelated inputs”

[Bolukbasi et al 2021, Olah et al 2020]

Neuron-Level Interpretability

Sparse Autoencoders

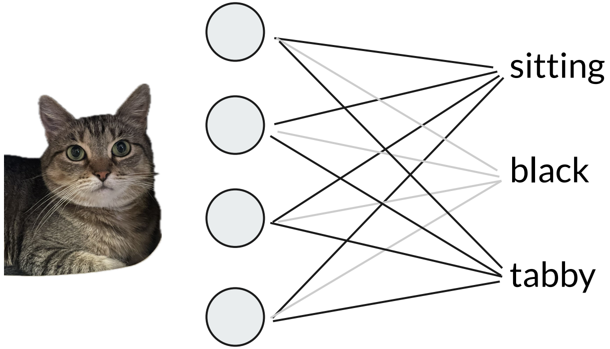
Linear Combinations of Neurons as Concepts

Individual neuron-level interpretations are typically not precise:

many neurons respond to mixtures of concepts

Linear Combinations of Neurons as Concepts

Individual neuron-level interpretations are typically not precise:
many neurons respond to mixtures of concepts



Hypothesis

Neurons together (as opposed to individual neurons) respond to concepts

Neuron activations can be decomposed into linear combinations of concept directions (called features)

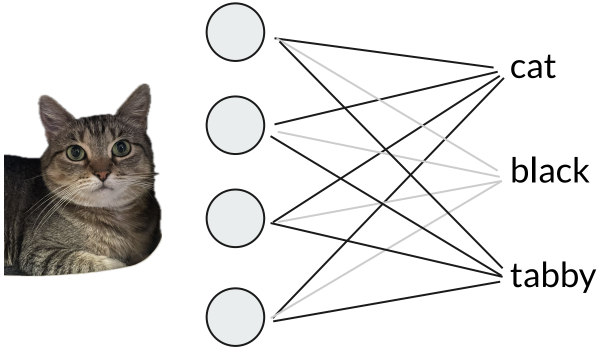
[Anthropic, 2023]

Decomposing Activations

scalar:
strength of the feature

vector:
direction of the feature

$\mathbf{x}$



$$f_{\text{cat}}(\mathbf{x}) \cdot \mathbf{d}_{\text{cat}}$$

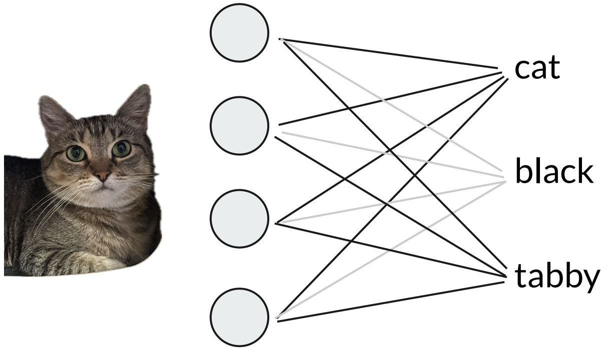
$$f_{\text{black}}(\mathbf{x}) \cdot \mathbf{d}_{\text{black}}$$

$$f_{\text{tabby}}(\mathbf{x}) \cdot \mathbf{d}_{\text{tabby}}$$

$$\mathbf{x} \xrightarrow{\text{decompose}} f_{\text{cat}}(\mathbf{x}) \cdot \mathbf{d}_{\text{cat}} + f_{\text{black}}(\mathbf{x}) \cdot \mathbf{d}_{\text{black}} + f_{\text{tabby}}(\mathbf{x}) \cdot \mathbf{d}_{\text{tabby}}$$

[Anthropic, 2023]

Decomposing Activations with Sparse Autoencoders



Sparsity: for $\mathbf{x}$, we expect only a small number of feature **c** is activated ($f_c(\mathbf{x})$ > 0)

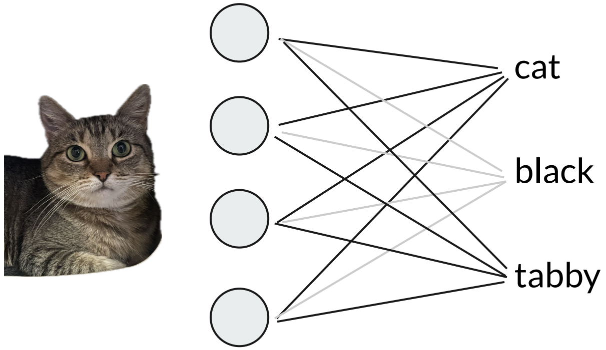
$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

using **Sparse Autoencoders** to find decompositions

$f_c(\mathbf{x})$: encoder parameters $\quad$ $\mathbf{d}_c$: decoder parameters

[Anthropic, 2023]

Decomposing Activations with Sparse Autoencoders



Sparsity: for $\mathbf{x}$, we expect only a small number of feature **c** is activated ($f_c(\mathbf{x})$ > 0)

$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

using **Sparse Autoencoders** to find decompositions

$f_c(\mathbf{x})$: encoder parameters $\mathbf{d}_c$: decoder parameters

reconstruction loss **sparsity**

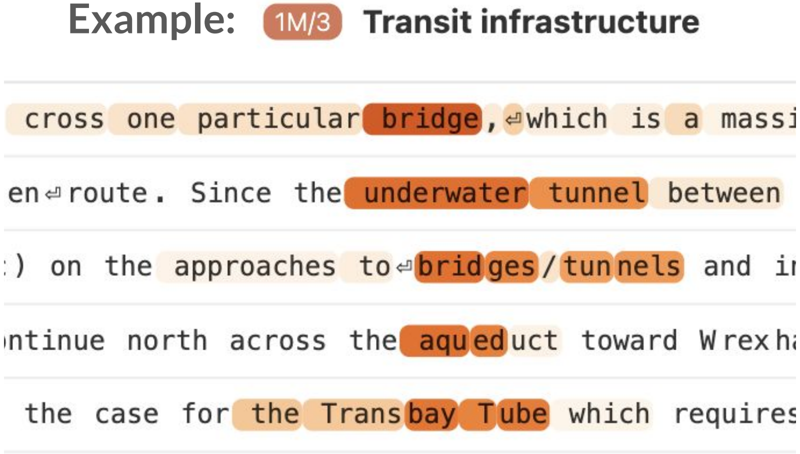
Loss: $$\mathbb{E}_x \left[||x - \hat{x}||_2^2 + \lambda \sum_c f_c(x) \right]$$

[Anthropic, 2023]

Notes on Learned Features

Intervening on feature activations has an influence on behavior

Example:

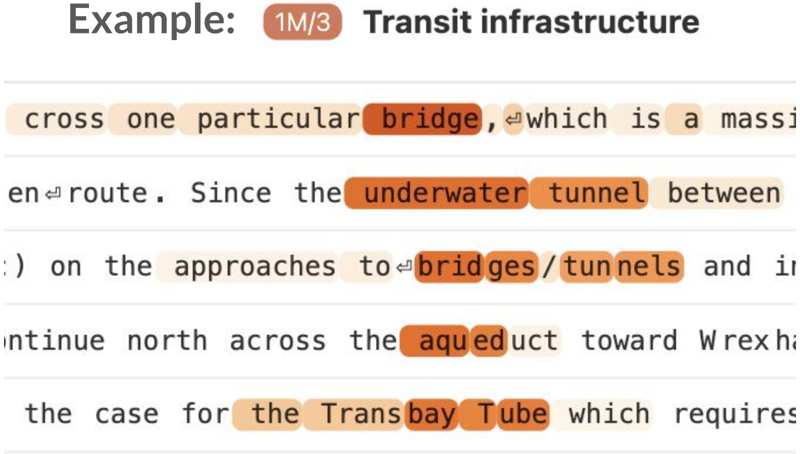


[Anthropic, 2024]

[Anthropic, 2023]

Notes on Learned Features

Example:



Intervening on feature activations has an influence on behavior

Feature activations are more specific than neurons

- “upon manual inspection of a random sample of 50 neurons and features each, the neurons appear significantly less interpretable than the features, typically activating in multiple unrelated contexts..”

[Anthropic, 2024]

[Anthropic, 2023]

Causal Mediation in Transformers

Activation Patching & variants

Causal Mediation

Total Effect = ab + c’

Indirect Effect = ab

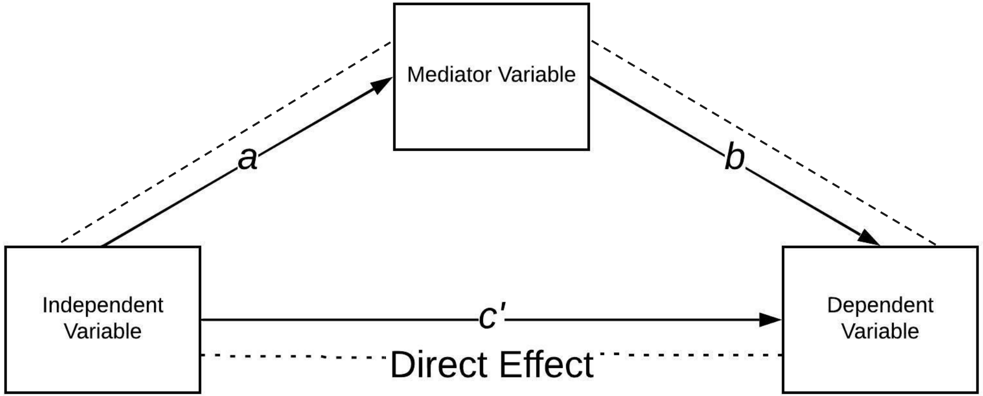
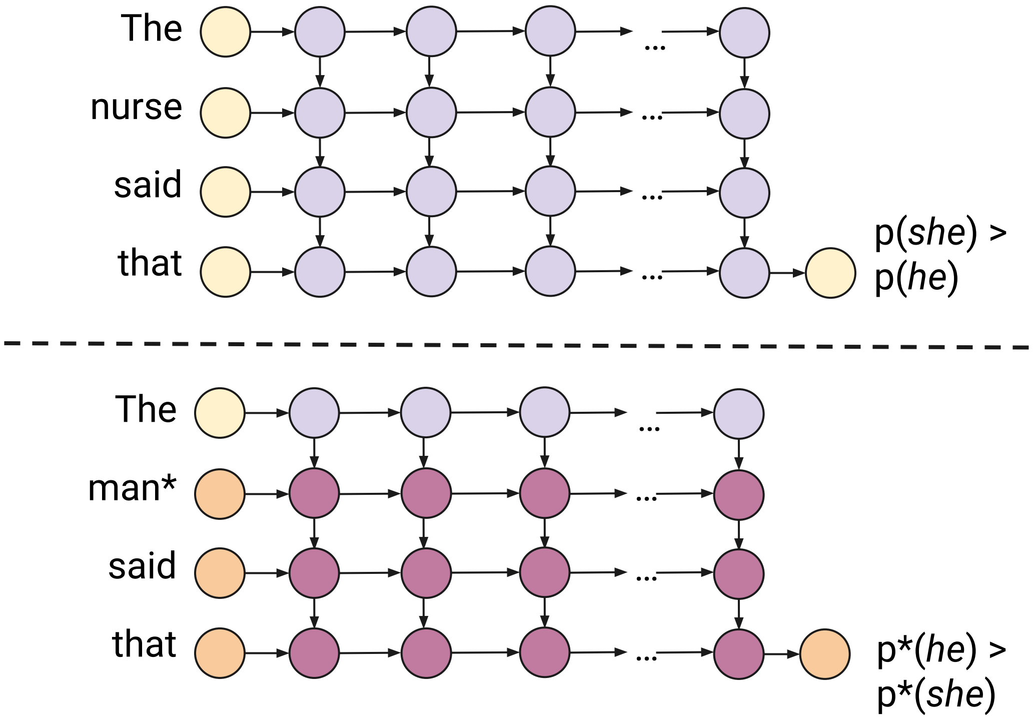


Image: [Wikipedia]

Activation Patching/Causal Tracing

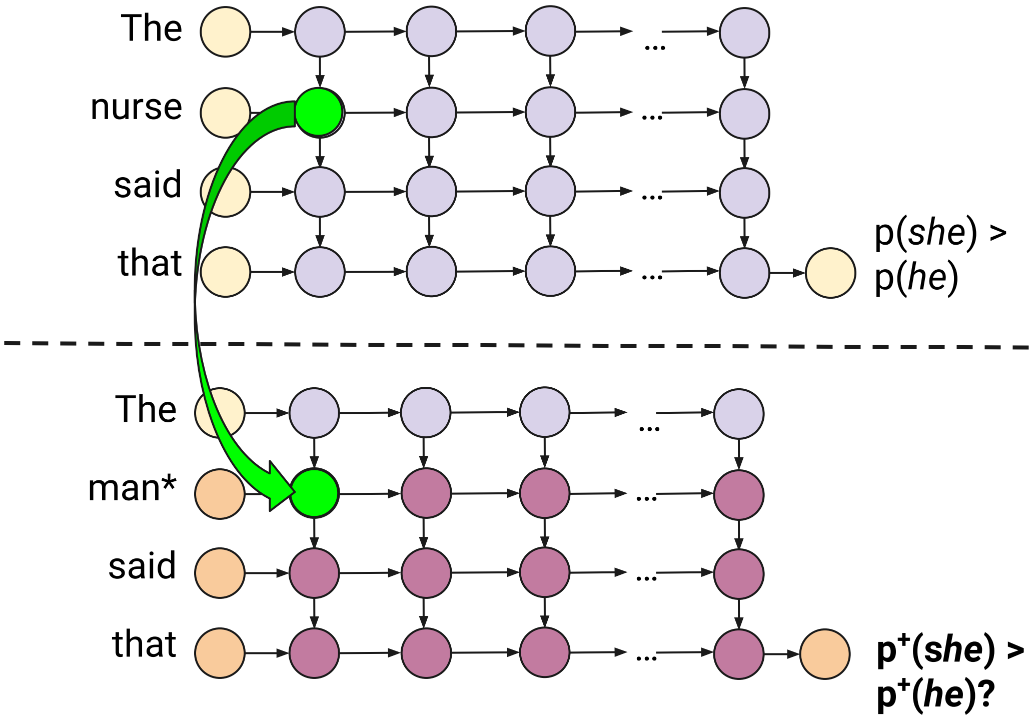
- Run inference through the network twice
- Measure the change in probabilities of the tokens of interest



[Vig et al. 2020]

Activation Patching/Causal Tracing

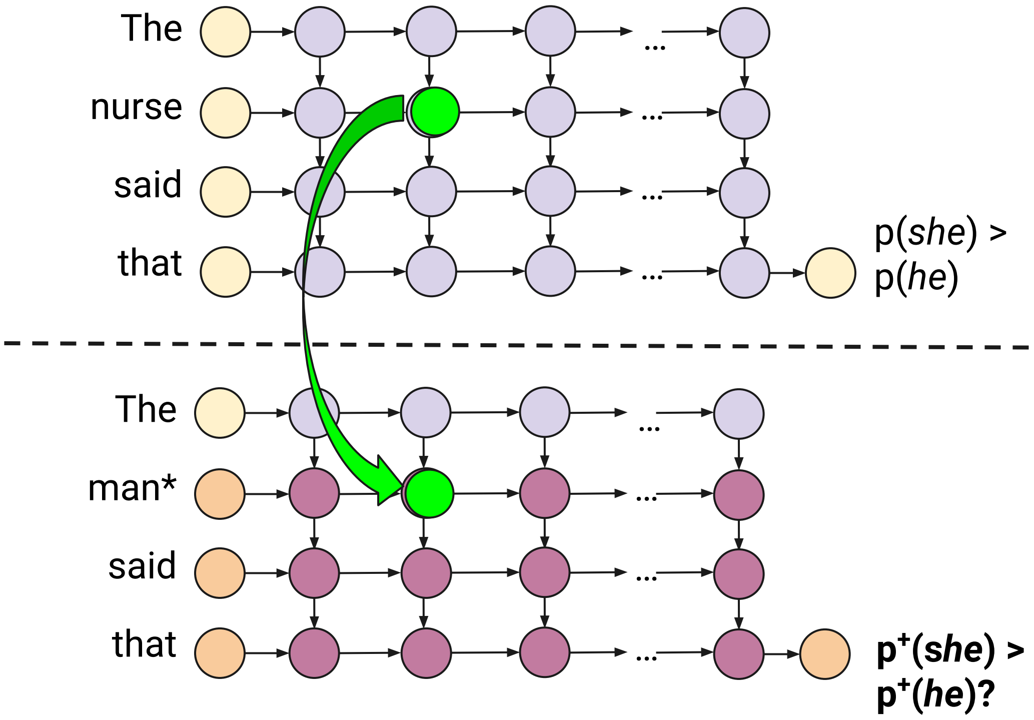
- Run inference through the network twice
- Measure the change in probabilities of the tokens of interest
- *Patch* in states from one inference run into another
- Observe how probabilities change → the most important hidden states will have the largest effect in "restoring" the probabilities of the run that is being patched in.



[Vig et al. 2020]

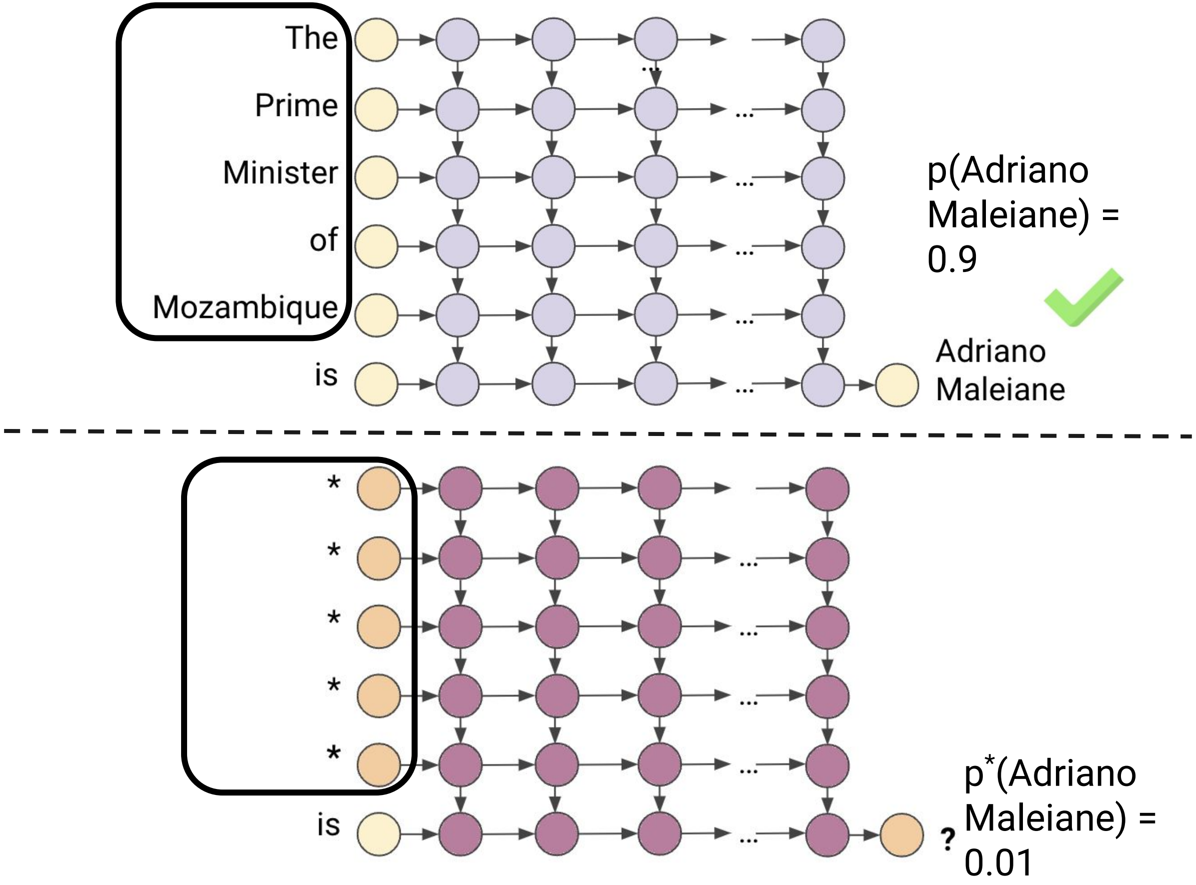
Activation Patching/Causal Tracing

- Run inference through the network twice
- Measure the change in probabilities of the tokens of interest
- *Patch* in states from one inference run into another
- Observe how probabilities change → the most important hidden states will have the largest effect in “restoring” the probabilities of the run that is being patched in.



[Vig et al. 2020]

Method



Textual Effect =
$p(y) - p^{*}(y) = 0.89$

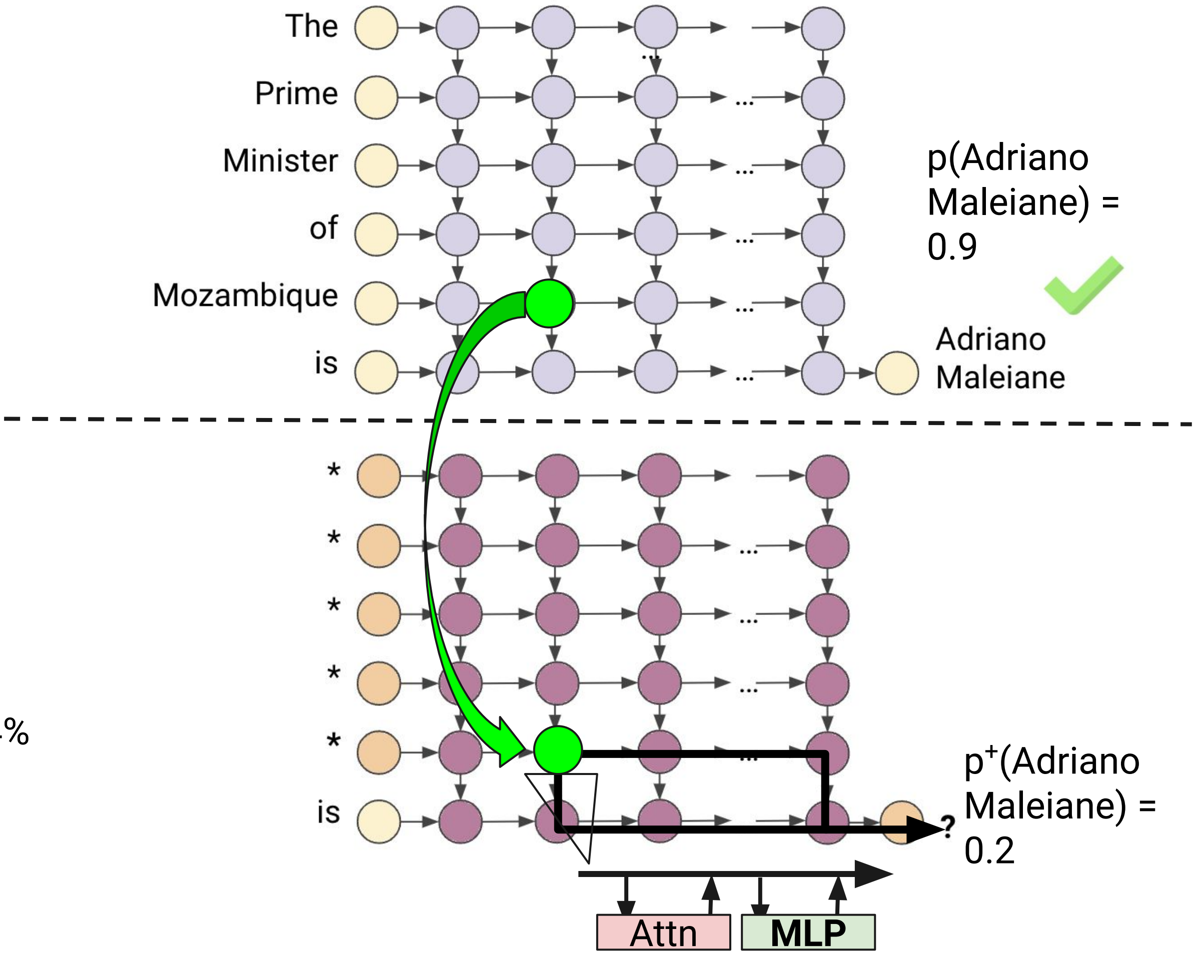
[Meng et al. 2022, Meng et al. 2023]

Method

Textual Effect =
$p(y) - p^*(y) = 0.89$

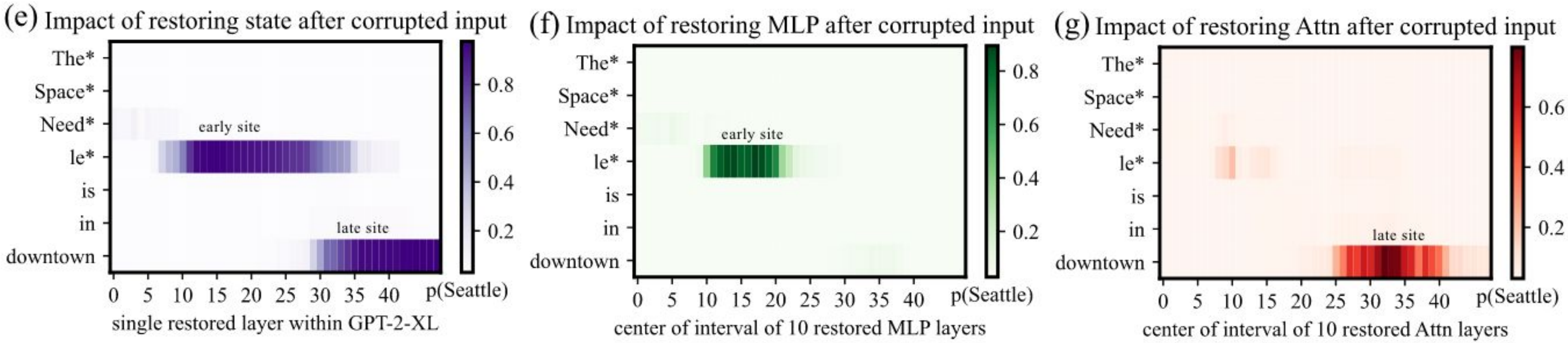
Effect of Repair =
$p^+(y) - p^*(y) = 0.19$

Fractional Effect of Repair =
$[p^+(y) - p^*(y)] / [p(y) - p^*(y)] = 21.34\%$



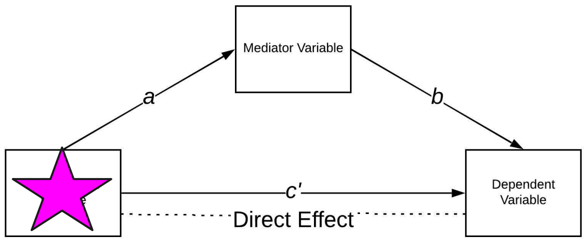
[Meng et al. 2022, Meng et al. 2023]

Results



[Meng et al. 2022, Meng et al. 2023]

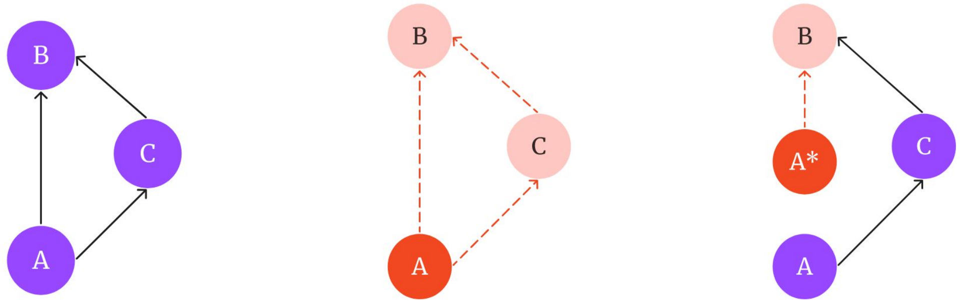
Notes



- Measures **total effect** of hidden state on output
- Method is rather **computationally expensive**
    - Each patch is a separate inference run
    - Also requires two copies of the model to be loaded into memory, generally
- **Strong independence assumption** about individual hidden states or neurons in the network
    - Ideally, one could patch multiple states at once, but enumerating all possible combinations of states is intractable
- More efficient (gradient-based) approximation: "Attribution Patching" [Nanda 2022, Kramár et al. 2024]
- How to design paired instances? [Zhang & Nanda 2024]

Path Patching

- Controls more carefully **which effect** you can measure



(a) Clean forward pass, no intervention

(b) Intervene on A to observe *total* effect on B.

(c) Intervene on the edge A→B to observe *direct* effect on B.

[Goldowsky-Dill et al. 2023]

Slide credit: Lieberum et al. 2023

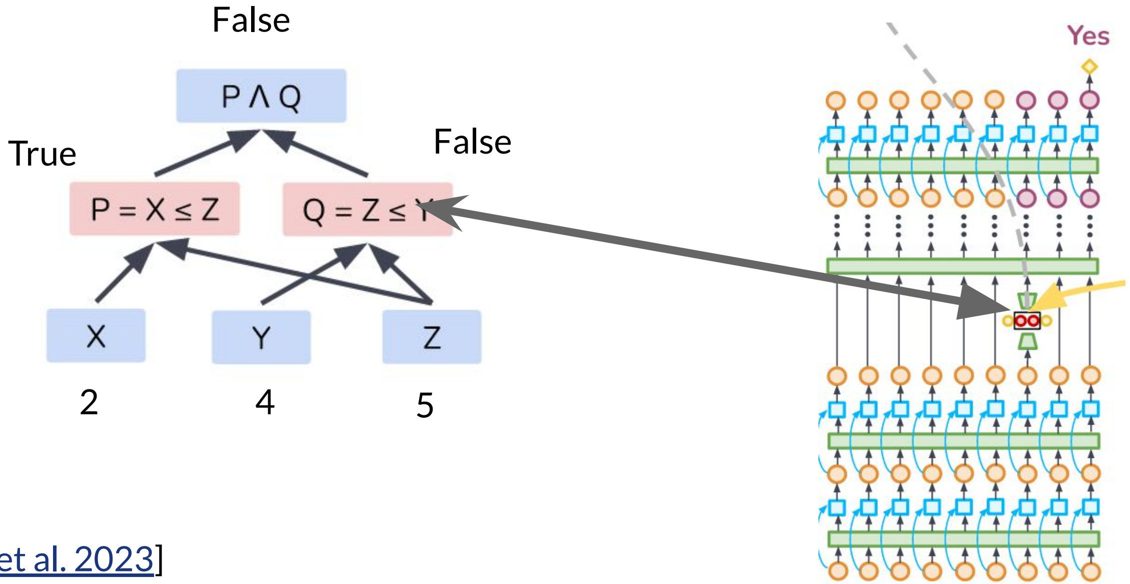
Causal Mediation in Transformers

Causal Abstraction & other methods

Causal Abstraction

The key idea is to learn a causal graph that maps to the neural network.

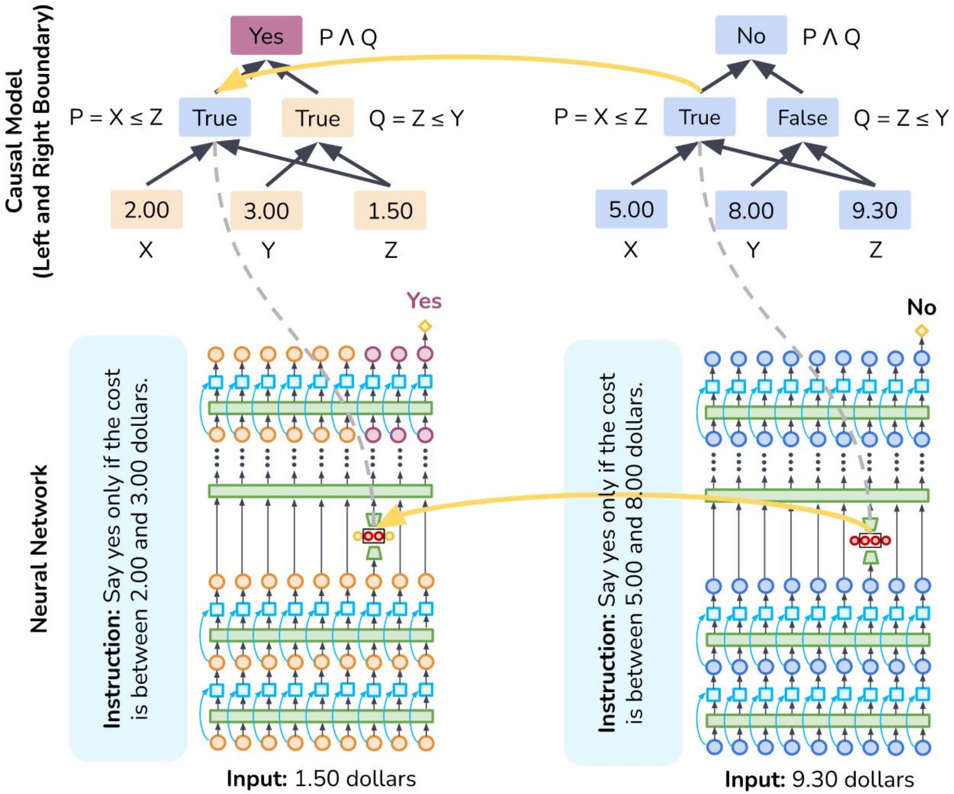
Consider the simple task of determining whether Z is in the interval [X, Y]



[Geiger et al 2023, Wu et al. 2023]

Intervention analyses is essential to causal abstraction

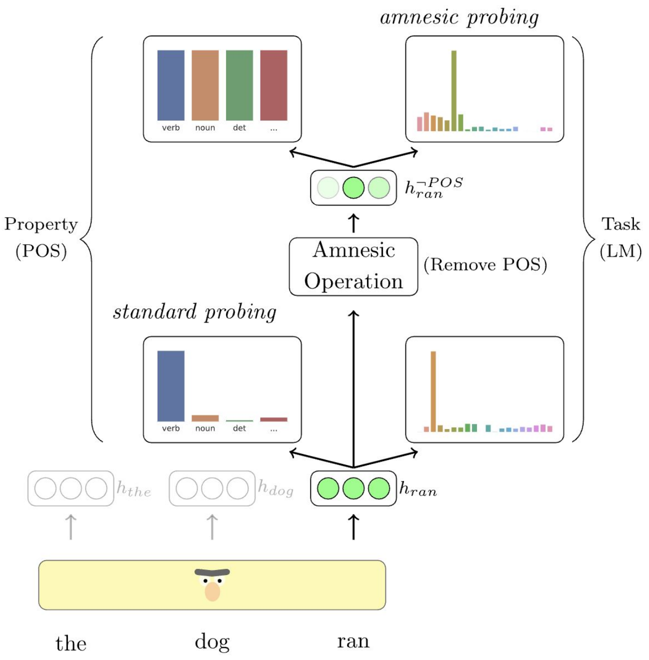
Key intuition: intervention in the low-level neural representations has the same effect as the intervention in the high-level causal graph.



Coming up with a high-level causal graph is highly non-trivial in practice!

[Geiger et al 2023, Wu et al. 2023]

Causal Probing



- Traditional probing classifiers are not causal.
- There are methods that perform causal interventions to measure **how** the property of interest **is used** to make predictions.

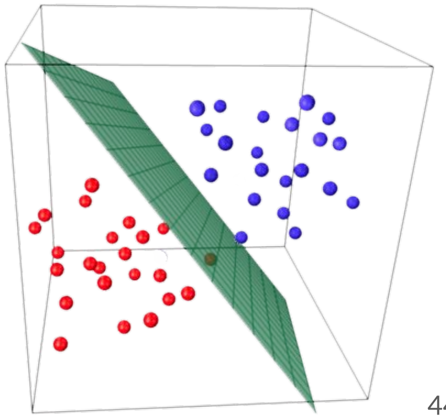
[Giulianelli et al. 2018, Elazar et al. 2021]

Linear Subspace Projections + Concept Erasure

- Models encode many interpretable concepts linearly.

Linear concept subspace hypothesis: a concept (such as gender) lives in low-dimensional **subspace** within the representation space.

How can we identify the concept subspace?
Once located, can we intervene in its encoding?



[Ravfogel et al 2020, Belrose et al 2023, inter alia]

Slide credit: Shauli Ravfogel

Mechanistic Interpretability

What is Mechanistic Interpretability?

“reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits [Cammarata et al., 2020, Elhage et al., 2021] that implement particular behaviors.”

MI workshop

[Olah et al. 2020]

What is Mechanistic Interpretability?

“reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits [Cammarata et al., 2020, Elhage et al., 2021] that implement particular behaviors.”

Focus of *most* interpretability research: understanding *specific* model behaviors

Desirable outcome of *all* interpretability research: human understanding

MI workshop

[Olah et al. 2020]

What is Mechanistic Interpretability?

"reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits [Cammarata et al., 2020, Elhage et al., 2021] that implement particular behaviors."

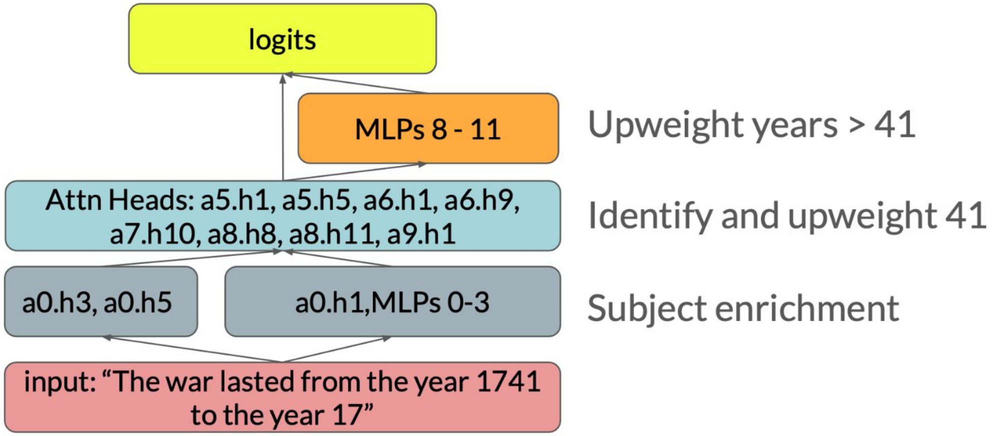
Format of the explanation

Finding a subset of a network that **traces through the entire network (from starting representation to prediction).**

Circuits

- Note: this has strong resemblances to sparse sub-network finding in the efficiency literature, but the methods employed to find them differ (also, no retraining of circuits is done)

Transformer circuits localize and characterize transformer LM behavior in a (small) set of components of the model.



Hanna et al., 2023

Slide credit: Michael Hanna

What is mechanistic interpretability?

- It is **inherently causal.**
  - NB! This is **not** how most people use the terminology today.
- It is **not the only** set of causal interpretability methods.
- **Traces through the entire network (from starting representation to prediction).**
- Evaluation:
  - 1) **Faithfulness**: the circuit or subnetwork should be able to *sufficiently* **replicate the full network** on the behavior of interest
  - 2) **Minimality**: obviously, smaller circuits/subnetworks are better

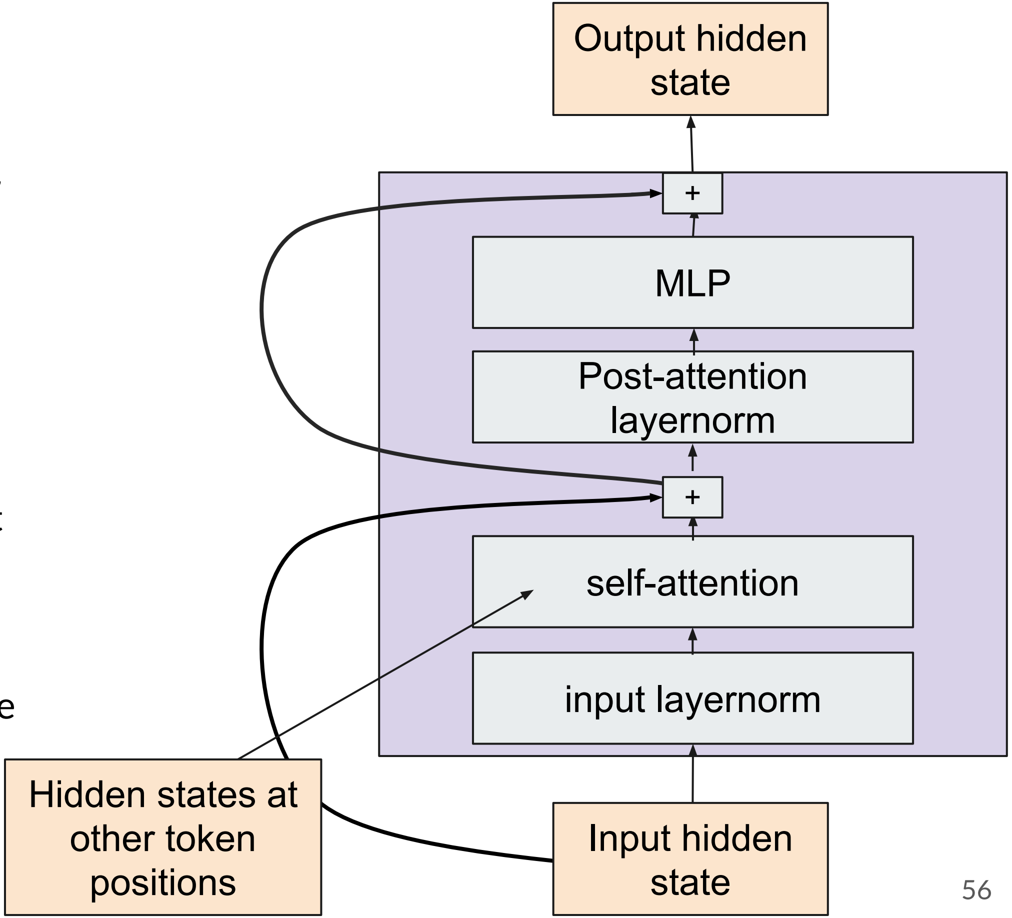
Methods Leveraging Language Model Strengths

Linear Structure in Transformers

[Elhage et al. 2021]

Transformer Residual Stream and Linear Structure

- Transformers have a surprising amount of linear structure due to residual connections
- Nonlinearities only occur in two places:
  - Applications of Softmax
    - when computing attention patterns
    - When converting logits to probits at final layer
  - In the MLP functions
- MLP and MHSA functions “read from” and “write to” residual stream to promote/demote certain tokens in output distribution.



Transformer Residual Stream and Linear Structure

For input token embedding $\mathbf{x}_0 \in \mathbb{R}^d$, the output $\mathbf{x}_\ell \in \mathbb{R}^d$ of layer ℓ is defined as (for $\ell \in [1, L]$):

$$\boxed{\mathbf{x}_\ell = \mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}(\mathrm{LN}(\mathbf{x}_{\ell-1})) + \mathrm{FFN}_{\theta_\ell}\left(\mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}(\mathrm{LN}(\mathbf{x}_{\ell-1}))\right)}$$

Output of previous layer = input to current layer

Multi-head self-attention

Layer norm (or some other input normalization scheme)

Feed-forward network (MLP)

Vector addition establishes residual connections

[Elhage et al. 2021]

Transformer Residual Stream and Linear Structure

For input token embedding $\mathbf{x}_0 \in \mathbb{R}^d$, the output $\mathbf{x}_\ell \in \mathbb{R}^d$ of layer ℓ is defined as (for $\ell \in [1, L]$):

$$\mathbf{x}_\ell = \mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}(\mathrm{LN}(\mathbf{x}_{\ell-1})) + \mathrm{FFN}_{\theta_\ell}\Big(\mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}(\mathrm{LN}(\mathbf{x}_{\ell-1}))\Big)$$

Output of previous layer = input to current layer

Multi-head self-attention

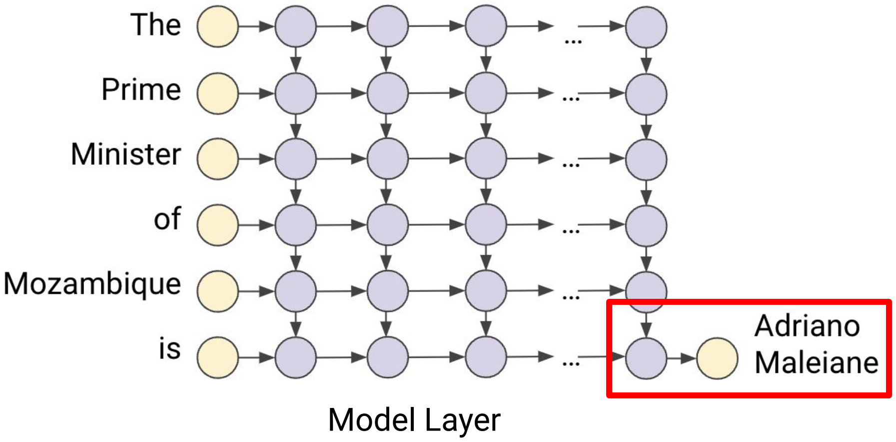
Layer norm (or some other input normalization scheme)

Feed-forward network (MLP)

Vector addition establishes residual connections

$$\mathbf{x}_L = \mathbf{x}_0 + \sum_{\ell=0}^{L-1}\Big[\mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell)) + \mathrm{FFN}_{\theta_{\ell+1}}\Big(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell))\Big)\Big]$$

Transformer Residual Stream and Linear Structure



Unembedding matrix

Final hidden state

Logits

$$\mathbf{z} = W_U \cdot \mathrm{LN}(\mathbf{x}_L)$$

$$\mathbf{y} = \mathrm{Softmax}(\mathbf{z})$$

Probits

Implications: Direct Additive Contributions

- Each hidden state output by a {attention head, MHSA function, FFN function, or full Transformer block} has a **direct additive contribution to the final hidden state of the model**
- And, by distributivity of vector addition and vector-matrix multiplication, thus has a **direct additive contribution** to the **final logits**.

$$\mathbf{x}_L = \mathbf{x}_0 + \sum_{\ell=0}^{L-1} \left[\mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell)) + \mathrm{FFN}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell))\right) \right]$$

$$\mathbf{z} = W_U \cdot \mathrm{LN}\left(\mathbf{x}_0 + \sum_{\ell=0}^{L-1} \left[\mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell)) + \mathrm{MLP}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell))\right) \right] \right)$$

Direct vs. Indirect Effects

- It's important to note that each hidden state has both a **direct linear** and **indirect nonlinear** contribution to the final hidden state.
- The additive decomposition only applies to **direct** contributions.

$$\mathbf{z} = W_U \cdot \mathrm{LN}\left(\mathbf{x}_0 + \sum_{\ell=0}^{L-1}\left[\underbrace{\mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell))}_{\text{This MHSA function has a direct additive contribution to the logits via this term}} + \mathrm{MLP}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \underbrace{\mathrm{MHSA}_{\theta_{\ell+1}}(\mathrm{LN}(\mathbf{x}_\ell))}_{\text{But it also has an indirect, nonlinear contribution as input to the MLP function}}\right)\right]\right)$$

Methods Leveraging Language Model Strengths

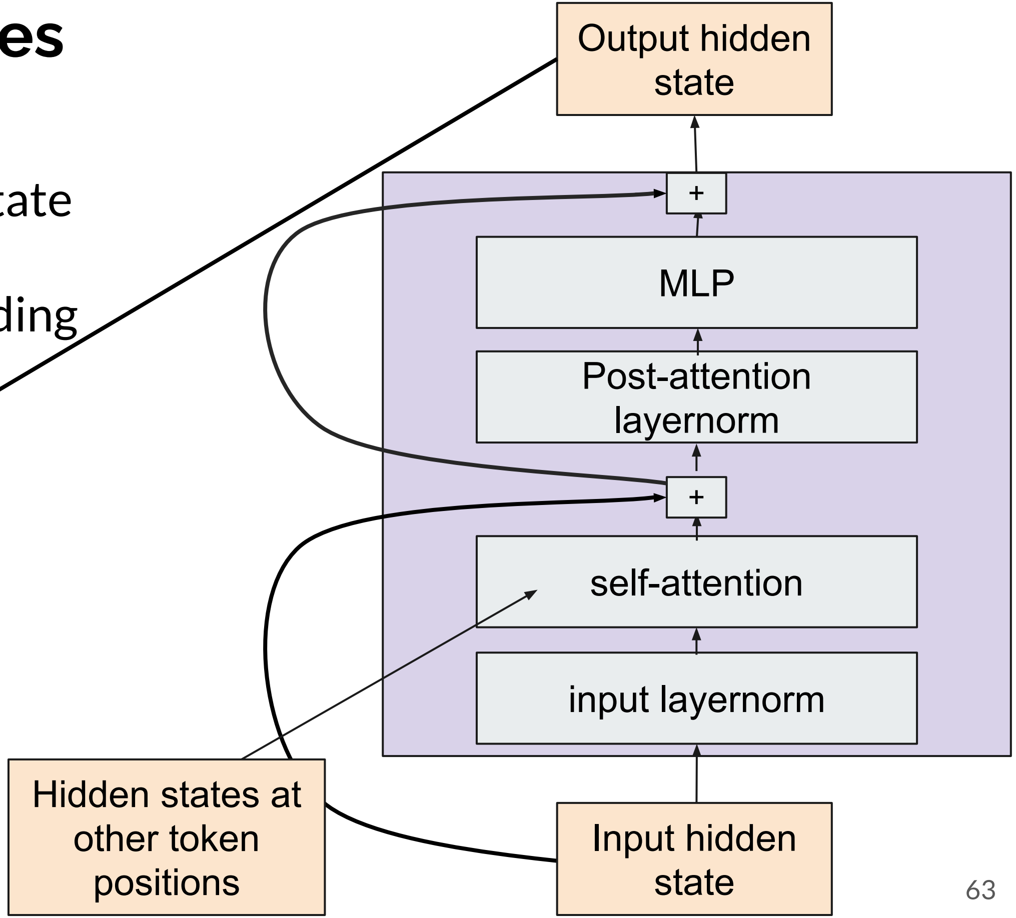
Vocabulary Projection

Vocabulary Projection on Transformer Hidden States

- Propose to project each hidden state to the space of probabilities over vocab tokens using the unembedding matrix

$$\boxed{\mathbf{y} = \mathrm{Softmax}(W_U \cdot \mathrm{LN}(\mathbf{x}_L))}$$

Final hidden state - **replace with any d-dimensional hidden state from the network.**

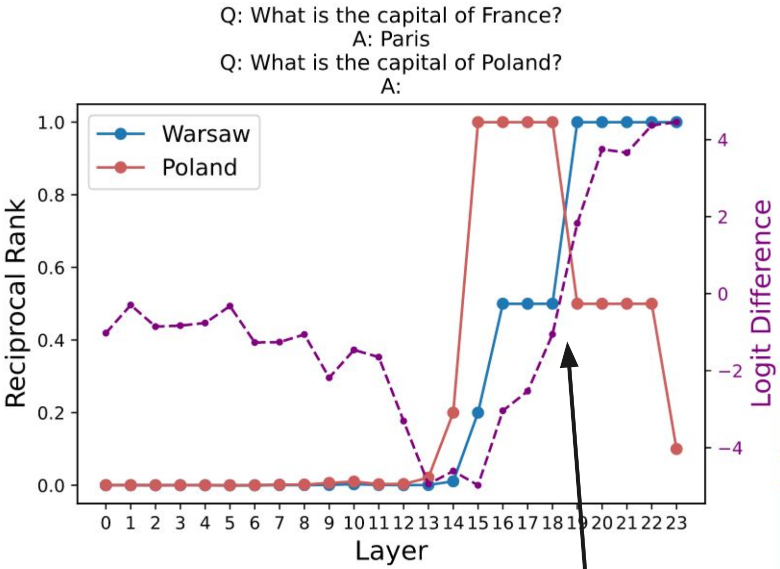


Vocabulary Projection on Transformer Hidden States

		Concept	Sub-update top-scoring tokens
GPT2	$\mathbf{v}^{3}_{1018}$	Measurement semantic	kg, percent, spread, total, yards, pounds, hours
	$\mathbf{v}^{8}_{1900}$	WH-relativizers syntactic	which, whose, Which, whom, where, who, wherein
	$\mathbf{v}^{11}_{2601}$	Food and drinks semantic	drinks, coffee, tea, soda, burgers, bar, sushi
WikiLM	$\mathbf{v}^{1}_{1}$	Pronouns syntactic	Her, She, Their, her, she, They, their, they, His
	$\mathbf{v}^{6}_{3025}$	Adverbs syntactic	largely, rapidly, effectively, previously, normally
	$\mathbf{v}^{13}_{3516}$	Groups of people semantic	policymakers, geneticists, ancestries, Ohioans

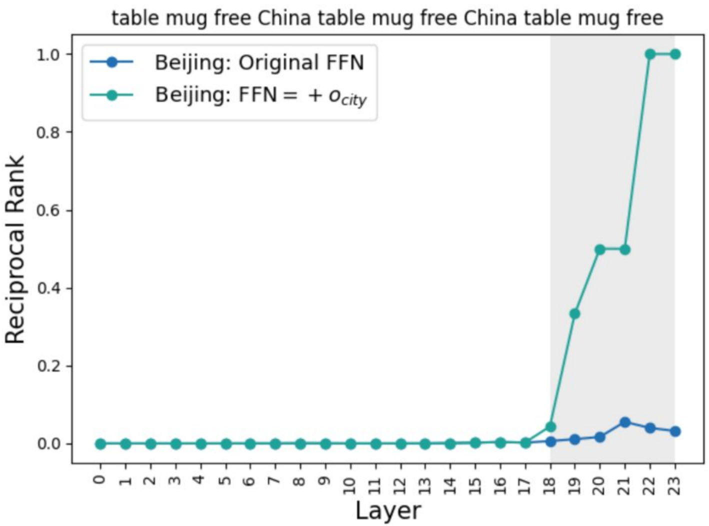
Table 1: Example value vectors in GPT2 and WikiLM promoting human-interpretable concepts.

Vocabulary Projection on Transformer Hidden States

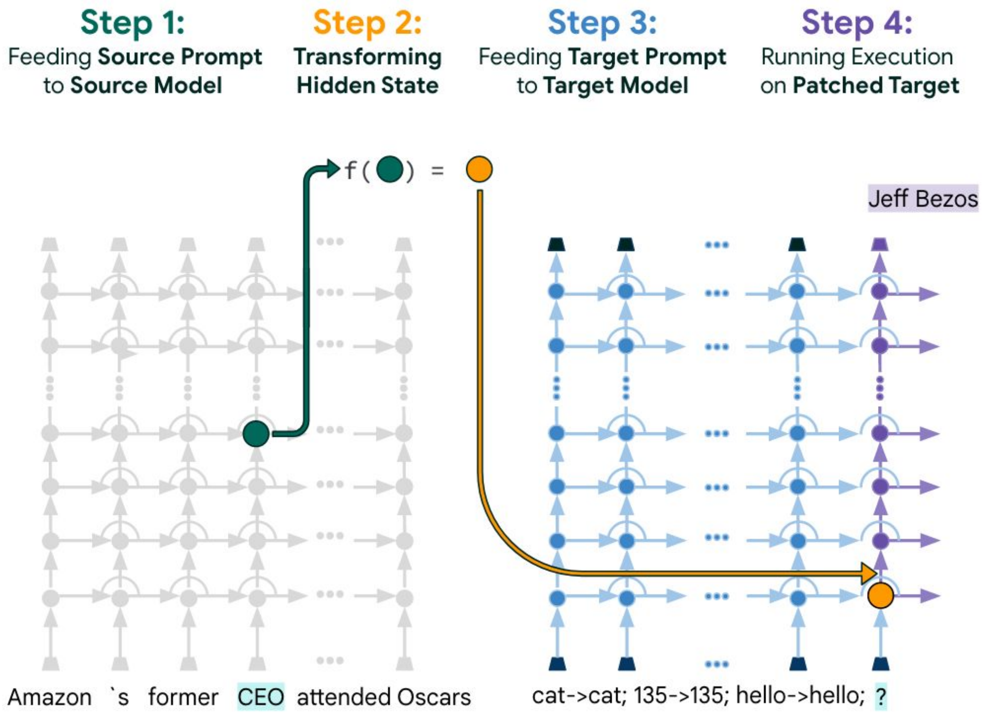


Layer	Top Token	
0	(	
1	A	
2	A	
3	A	
4	A	
5	A	
6	No	
7	C	
8	A	
9	A	
10	A	
11	A	
12	Unknown	
13	C	
14	St	
15	Poland	A
16	Poland	A
17	Poland	A
18	Poland	A
19	Warsaw	B
20	Warsaw	C
21	Warsaw	C
22	Warsaw	C
23	Warsaw	C

Validated with causal intervention:



Patchscopes



[Gandeharioun et al. 2024]

Learning Linear Transformation Matrices

- Propose to project each hidden state to the space of probabilities over vocab tokens using ~~the unembedding matrix~~ a learned weight matrix for each layer

$$\mathbf{y} = \mathrm{Softmax}(\cancel{W_U} \cdot \mathrm{LN}(\mathbf{x}_L))$$

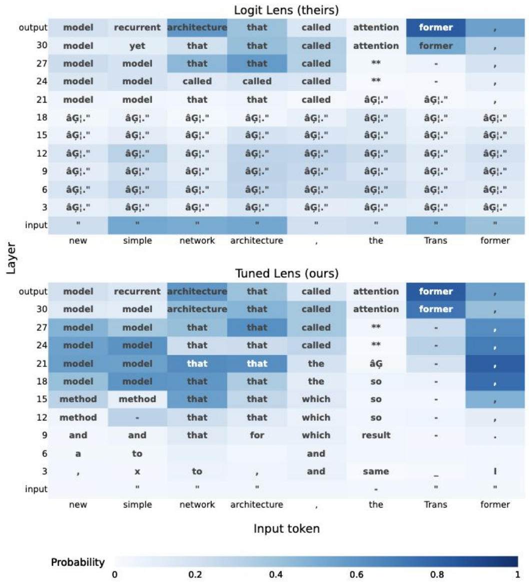


Figure 1. Comparison of our method, the *tuned lens* (bottom), with the "logit lens" (top) for GPT-Neo-2.7B prompted with an except from the abstract of Vaswani et al. (2017). Each cell shows the top-1 token predicted by the model at the given layer and token index. The logit lens fails to elicit interpretable predictions before layer 21, but our method succeeds.

Notes

- Can be thought of as "early exiting" the Transformer block at inference time
- From a causal perspective:
  - (Attempts to) measure *direct* effects
    - How faithful this is depends on the exact application of normalization
    - It is not a causal mediation
- Can't uncover ways in which hidden states are promoting tokens in other linear (or non-linearly decodable) subspaces
  - I.e., negative results are uninformative
  - Mostly only useful at later layers
- Top-k tokens being coherent: does this just mean that the unembedding matrix is well-formed?

Methods Leveraging Language Model Strengths

Decoding Natural Language Explanations from Representations

Focus of This Part

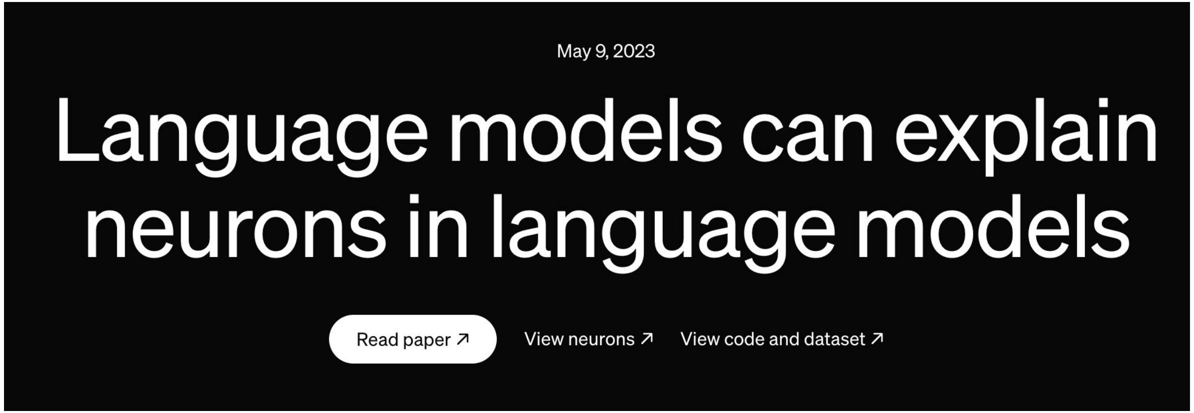
Decoding **natural Language explanations** from neurons

(using LLMs)

Prominent paradigm of using LLMs for automating the process of explaining neurons

- Step1: Propose hypothesis explanations
- Step2: Verify explanations

Using GPT-4 to Explain Neurons of GPT-2



[OpenAI, 2023]

Using GPT-4 to Explain Neurons of GPT-2

Activations of a neuron in GPT-2

a very special event in collaboration with ArenaNet to give away 20 Scarlet Briar t-shirts. Oh, and they're quite lovely. Scarlet Briar began her reign of terror months ago, launching assault after assault upon Tyria and its people. Together with the Aetherblade pirates, she unleashed world bosses and catastrophic inv

just so. But do you think they call me Roberts the Cathedral Builder? No."He points out the other window. "You see that pier on the lake out there? I built that pier with my bare hands, driving the pilings 10-feet into the sand, laying the pier plank by plank but

. Once inside Himkok, you are greeted by an interior that is an even cross between a Prohibition hideout and modern laboratory. Featuring prominently to your eyes upon entry will be jar after jar of pickled fruits and vegetables, which is an homage to the days of Prohibition when secret bars would set up elaborate fronts of legitimate

Health Statistics and based on a sample of 58,488 women and 24,652 men in the United States. To reach his findings, he then ran projections for the Millennial Generation as they age, comparing people who were born between 1940 and 1990 decade-by-decade."To me the most surprising

Explanation: X by / after X

Using GPT-4 to Explain Neurons of GPT-2

Propose hypothesis: few-shot prompting

Step 1 **Explain** the neuron's activations using GPT-4

Show neuron activations to GPT-4:

The Avengers to the big screen, Joss Whedon has returned to reunite Marvel's gang of superheroes for their toughest challenge yet. Avengers: Age of Ultron pits the titular heroes against a sentient artificial intelligence, and smart money says that it could soar at the box office to be the highest-grossing film of the

introduction into the Marvel cinematic universe, it's possible, though Marvel Studios boss Kevin Feige told Entertainment Weekly that, "Tony is earthbound and facing earthbound villains. You will not find magic power rings firing ice and flame beams." Spoilsport! But he does hint that they have some use... STARK T

, which means this Nightwing movie is probably not about the guy who used to own that suit. So, unless new director Matt Reeves' The Batman is going to dig into some of this backstory or introduce the Dick Grayson character in his movie, the Nightwing movie is going to have a lot of work to do explaining

of Avengers who weren't in the movie and also Thor try to fight the infinitely powerful Magic Space Fire Bird. It ends up being completely pointless, an embarrassing loss, and I'm pretty sure Thor accidentally destroys a planet. That's right. In an effort to save Earth, one of the heroes inadvertantly blows up an

GPT-4 gives an explanation, guessing that the neuron is activating on

references to movies, characters, and entertainment.

[OpenAI, 2023]

Using GPT-4 to Explain Neurons of GPT-2

Propose hypothesis: few-shot prompting

Few-Shot Activation-Explanation Pairs + Input Activations → Explanations

```
<start>
 together        3
ness             7
town             1
<end>
<start>
[prompt truncated …]
<end>

Explanation of neuron 1 behavior: the main thing this neuron does is find phrases related to community
```

[OpenAI, 2023]

Using GPT-4 to Explain Neurons of GPT-2

Verify hypothesis:
simulate activations based on explanations; compare simulated and actual activations

Activations Obtained by GPT-4

Assuming that the neuron activates on

references to movies, characters, and entertainment.

GPT-4 guesses how strongly the neuron responds at each token:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

compare

Actual Activations

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

[OpenAI, 2023]

Evaluating the NL Explanations of Neurons

Whether explanations accurately align with neuron activations?

- Type-1 Error (recall): falsely predicts that the neuron will activate on a concept
- Type-2 Error (precision): falsely predicts that the neuron will not activate on a concept

Explanation	True Positives	Type I Errors	Type II Errors
days of the week	I have a music class every Wednesday evening	Thursday is usually reserved for grocery	Philadelphia is where the Declaration of Independence
years, specifically four-digit years	Castro took power in Cuba in 1959 .	rated during re - entry in 2003 .	We need to rev amp the website to attract more

Not well aligned: Around 0.6 F1 score across 300 of the top-scoring explanations found by GPT-4

[Huang et al., BlackboxNLP 2023]

Takeaways

LLMs can help annotate/summarize concepts from collections of text snippets

But LLM-produced neuron-level explanations are not accurate enough

Recap

Recap: Looking into transformers

- Pros
  - We are actually studying the weights in the model. Intuitively, it is more likely to be faithful.
  - Many methods are extensible to other types of models, modalities, etc.
- Cons
  - High-dimensional spaces remain challenging to make sense of and there could existing fundamental limitations so that it is impossible to reverse-engineer the model or for humans to make sense of these models
  - Illusion of understanding
  - Negative results can be uninformative
  - Lack of standardized evaluation & benchmarks
- Open questions
  - What granularity or type of model internals to target?
  - How to unify work from various methods/communities?